6

Choices of Expression

		Page
6A.	One-way comparisons	6-1
6B.	How should we look?	6-9
6C.	Some two-way examples	6-12
6D.	Looking at the evidence	6-24
6E.	Stripes and skews	6-35
6F.	How far have we come?	6-48

*****

6R.	Why do we look at the evidence as we do?	6-51
6S.	Matched powers, how do they differ?	6-54
6T.	Why is (row adjustment)x(column adjustment) its own residual?	6-62

2I Location of tables and plots for 6

Number	Page	Number	Page
1	6-2	26	6-44
2	6-3	27	6-46
3	6-5	28	6-55
4	6-7	29	6-58
5	6-8	30	6-58
6	6-10	31	6-60
7	6-13	32	6-61
8	6-14		
9	6-16		
10	6-17		
11	6-18		
12	6-19		
13	6-20		
14	6-22		
15	6-26		
16	6-27		
17	6-29		
18	6-31		
19	6-32		
20	6-33		
21	6-37		
22	6-38		
23	6-39		
24	6-41		
25	6-42		

6

Choices of Expression

In chapters 2 and 3, we gave considerable attention to how our data was expressed, and to better choices of how this was done. Chapters 4 and 5 thus became notable for saying nothing about choice of expression -- not because there was nothing to say, but rather because the analytic mechanism to be built up was broad, deep, and unfamiliar enough to deserve concentrated attention. It is now time to come to the important question of how to look at the results of the analysis and see whether the expression of the data deserves a change.

6A. One-way comparisons

In 1940, Winsor and Clarke reported the results of a study of the catch of plankton (small ocean life) in nets of different size and in nets of the same kind hauled successively or simultaneously. Table 1 shows a small part of their data. If we proceed to analyze this data as giving all comparisons of three net-depth combinations in each of nine blocks, we would find the analysis shown in table 2. So far as the size of the combining residuals goes, this analysis seems quite satisfactory. With differences as large as ±1800, the residuals are only ±7.

Let us go on. Table 3 continues the analysis, finding individual residuals. The evidence of something large and systematic going on is overwhelming. We need to find out:

- what is going on in this set of data.
- how to deal with it.

table 1 of chapter 6

Catch of Sagitta elegans by Plankton Nets
of 2 Diameters at 3 Depths Hauled Simultaneously
(Many catches estimated from samples.)

Haul	12.7 cm at 29 meters (A)	75 cm at 30 meters (B)	12.7 cm at 31 meters (C)
1	45	3020	7
2	100	4350	62
3	117	6060	76
4	110	6600	65
5	(one haul lost)		
6	27	1240	27
7	31	1240	52
8	20	1040	7
9	24	415	6
10	27	690	5

Source: C. P. Winsor and G. L. Clarke 1940. A statistical study of variation in the catch of plankton nets. Journal of Marine Research 3: 1-34, especially table VIII at page 33.

Data location: Cape Cod Bay, off Canal Entrance; 9 September 1937; 10 minute hauls at 2 knots.

table 2 of chapter 6

Naive Analysis of the Data of Table 1

a) THE DATA, REARRANGED FOR CLARITY

	Net & depth		
	C	A	B
4	65	110	6600
3	76	117	6060
2	62	100	4350
1	7	45	3020
6	27	27	1240
7	52	31	1240
8	7	20	1040
9	6	24	415
10	5	27	690

b) DIRECT DIFFERENCES, WITH EASY MIDMEANS

	A-C	B-A	B-C
4	45	6490	6535
3	41	5943	5984
2	38	4250	4288
1	38	2975	3013
6	0	1213	1213
7	-21	1209	1188
8	13	1020	1007
9	18	391	373
10	22	663	641
(*)	26	1800	1805

c) SQUARE COMBINING TABLE

Minus net & depth	Plus net & depth		
	A	B	C
		1800	-26
A	0		
B	-1800	0	-1805
C	26	1805	0
mean	-591	1202	-610

2I

table 2 of chapter 6 (Cont'd) - 2

d) FITTED DIFFERENCES

	A	B	C
A	0	1793	-19
B	-1793	0	-1812
C	19	1812	0

e) COMBINING RESIDUALS

A	0	7	-7
B	-7	0	7
C	7	-7	0

table 3 of chapter 6

Continuation of the Naive Analysis of table 2

a) ADJUSTED VALUES (Adjusted for net & depth)

	Net & depth		
Haul	C	A	B
4	675	701	5398
3	686	708	4858
2	672	691	3148
1	617	636	1818
6	637	618	38
7	662	622	38
8	617	611	-162
9	616	615	-787
10	615	618	-512

b) INDIVIDUAL RESIDUALS

4	-26	0	4697
3	-22	0	4150
2	-19	0	2457
1	-19	0	1182
6	19	0	-580
7	40	0	-580
8	6	0	-773
9	1	0	-1402
10	0	3	-1127

c) EXAMPLES. 675 = 65 (from 2(a)) - (-610) (from 2(c))

-26 = 675 (from (a)) - 701 (easy midmean for haul)

- how to look for similar effects when they do not shout at us as loudly as in this example.

What is going on is, I hope, clear to all of us who have stopped to look at table 1 and think about its numbers. The big net, six times the diameter of the small nets, would naively be expected to catch something like 36 times as much as the small nets. While the catch in the big net is falling by 5000, say from 6000 to 1000, we would expect a proportionate fall in the catch in the small nets, a fall by one or two hundred -- certainly not a fall by 5000 from 200 to an impossible figure of MINUS 4800. Differences in numbers caught are just not a sensible thing to look at. We should be looking at ratios. As usual, the easy way to look at ratios is to look at differences of logs.

Table 4 repeats the analysis of table 2 for logs. Again the combining residuals are pleasantly small.

What of the individual residuals? Table 5 continues the analysis. Now the residuals:

- lack the systematic appearance that shouted at us in the earlier analysis.
- show clearly, especially in the stem-and-leaf display, several far-out values.

The fact that the large residuals occur for the very first haul of one net and the last two hauls of another is suggestive. Whether this is associated with changes of Sagitta density or with human frailty in rigging nets, or what, we are not prepared to judge. Clearly, however, we would not have had these oddities brought to our attention had we confined ourselves to the analyses of tables 2 and 3.

table 4 of chapter 6

The Analysis of Table 2 Repeated For the Logs of the Catch

Haul	Net & depth		
	C	A	B
a) THE DATA			
4	1.81	2.04	3.82
3	1.88	2.07	3.78
2	1.79	2.00	3.64
1	.85	1.65	3.48
6	1.43	1.43	3.09
7	1.72	1.49	3.09
8	.85	1.30	3.02
9	.78	1.38	2.62
10	.70	1.43	2.84

b) DIRECT DIFFERENCES, WITH EASY MIDMEANS

	A-C	B-A	B-C
4	.23	1.78	2.01
3	.19	1.71	1.90
2	.21	1.64	1.85
1	.80	1.83	2.63
6	.00	1.66	1.66
7	-.23	1.60	1.37
8	.45	1.72	2.17
9	.60	1.24	1.84
10	.73	1.41	2.14
(*)	.30	1.67	1.92

Minus net & depth	Plus net & depth		
	A	B	C
c) SQUARE COMBINING TABLE			
A	0	1.67	-.30
B	-1.67	0	-1.92
C	.30	1.92	0
mean	-.46	1.20	-.74
d) FITTED DIFFERENCES			
A	0	1.66	-.28
B	-1.66	0	1.94
C	.28	1.94	0
e) COMBINING RESIDUALS			
A	0	.01	-.02
B	-.01	0	+.02
C	.02	-.02	0

table 5 of chapter 6

Continuation of the Analysis of the Logs

Haul	Net & depth C	A	B
a) ADJUSTED VALUES			
4	2.55	2.50	2.62
3	2.62	2.53	2.58
2	2.53	2.46	2.44
1	1.59	2.11	2.28
6	2.17	1.89	1.89
7	2.46	1.95	1.89
8	1.59	1.76	1.82
9	1.52	1.84	2.42
10	1.44	1.87	2.64
b) INDIVIDUAL RESIDUALS			
4	.00	-.05	.07
3	.04	-.05	.00
2	.07	.00	-.02
1	-.52	.00	.17
6	.28	.00	.00
7	.51	.00	-.06
8	-.17	.00	.06
9	-.32	.00	.58
10	-.43	.00	.77

c) STEM-AND-LEAF DISPLAY OF RESIDUALS

.7	7
.6	
.5	18
.4	
.3	
.2	8
.1	7
.0	740070060
-.0	0510520600
-.1	7
-.2	
-.3	2
-.4	3
-.5	2

Note: Allowance may need to be made for nine zeros forced by use of easy midmeans. Even if this allowance be made, however, +.77, +.58 and -.52 are "far out".

We have found the source of the systematic residuals in our first analysis, and have found a choice of expression that both eliminates them and allows a much more revealing analysis of what else is going on.

6B. How should we look?

The example of the last section involved a loud and blaring signal that choice of expression required attention. What if the strongest signal available is a loud whisper? What can we do to be able to hear it?

Table 6 goes through the analysis of a simple artificial example, where we can see what happens and which handles we have to hold onto. In view of the basic analysis, the squared data ought to show a need for a change of expression. The residuals, numerical or coded, of the squared values, are not revealing in their original array. When rearranged according to the orders of row adjustments and column adjustments, however, either coded or numerical residuals show forth a very distinctive pattern, one in which the NE and SW corners deviate in one direction while the NW and SE corners deviate in the other.

This pattern of residuals is very distinctive. It illustrates the main grasp we have on the need of a row-by-column table for a simple, better choice of expression.

For many situations, what we have learned is all we need. In others, it may help:

- to have a little guidance as to what is a "corner" of the table of residuals.
- to be able to grasp the handle more carefully by plotting the numerical residuals appropriately, when care is worthwhile.

table 6 of chapter 6

Elementary Analysis of an Artificial Example
According to Two Expressions

a) THE DATA

	Basic A	Basic B	Basic C	Squared A	Squared B	Squared C
1	5	7	6	25	49	36
2	2	4	3	4	16	9
3	1	3	2	1	9	4
4	4	6	5	16	36	25

b) ELEMENTARY ANALYSIS

1	0	0	0	2	-4	4	0	19
2	0	0	0	-1	2	-2	0	-8
3	0	0	0	-2	4	-4	0	-13
4	0	0	0	1	-2	2	0	8
	-1	1	0	4	-7	9	0	17

c) CODED RESIDUALS (AS IS) FOR SQUARED VALUES

		A	B	C	
1	1	L··	··H	·····	19
2	2	···	···	·····	-8
3	3	··H	L··	·····	-13
4	4	···	···	·····	8
	---	-7	9	0	17

d) CODED RESIDUALS REARRANGED

17	-7	0	9	...
19	L··	·····	··H	1
8	···	·····	···	4
-8	···	·····	···	2
-13	··H	·····	L··	3
	A	C	B	

e) NUMERICAL RESIDUALS REARRANGED

17	-7	0	9	...
19	-4	0	4	1
8	-2	0	2	4
-8	2	0	-2	2
-13	4	0	-4	3
	A	C	B	

- to know how set up the present analysis is so that it will apply to other kinds of partial description.

We shall try to provide answers to these questions at various points throughout the remainder of this chapter, looking first at an "easy" analysis based on division into thirds. This is rather crude, but often useful.

Suppose that, for a larger table, the column adjustments run as follows when put in order:

-18, -1, -1, -1, -1, 0, 0, +1, +1, +6, +10, +52

What columns are to contribute to the corners? The convenient rule is to take the interval from lowest to highest and divide this into three parts of equal length. Here the length is 70 and the parts run:

- from -18 to +5
- from +6 to +28
- from +29 to +52

Thus nine columns can contribute to the NW and SW corners, but only 1 to the NE and SE, leaving the two columns whose adjustments are +6 and +10 to define the middle third and contribute to no corner. (Clearly the middle third could have been empty.)

Rows should be divided into thirds in a similar way.

Even though this procedure can produce very unsymmetrical looking corners -- in a 12x12 table where both sets of stripes -- rows and columns -- broke as in our example, one corner would contain 81 residuals, while another would contain a lonely one -- it seems to be the simplest satisfactory choice for a standard "easy" inquiry.

6C. Some two-way examples

Elementary analysis of presidential voting in 10 Nebraska counties in the 12 presidential elections from 1920 to 1964 -- an example we will meet again -- gives the results shown in table 7. The column (election) adjustments range from -10.6 to +19.6, so that the thirds are:

-10.6 to -0.6

-0.5 to +9.5

+9.6 to +19.6

The row (county) adjustments range from -5.0 to +17.1, so that the thirds are:

-5.0 to +2.3

+2.4 to +9.7

+9.8 to +17.1

Table 8 shows the numerical residuals with rows and columns rearranged and the thirding lines drawn in. A casual glance shows only a mass of numbers. Looking a little closer we find that there was only one high county and no middle counties. In fact, the numbers of observations in each of the 9 = 3x3 cells generated by thirding are:

NW		NE
6	3	3
0	0	0
54	27	27
SW		SE

table 7 of chapter 6

Elementary Analysis of % Vote Democratic (of Major Party Vote)
in 12 Presidential Elections and 10 Counties of Nebraska

	'20	'24	'28	'32	'36	'40	'44	'48	'52	'56	'60	'64	
Adams	-4.2	-2.9	-3.4	-5.1	3.1	0.1	1.2	1.6	2.7	2.0	2.2	2.4	+0.2
Buffalo	0.8	-1.2	-3.5	-1.6	1.3	1.1	2.3	1.3	-2.3	0.1	2.1	0.4	-1.3
Dawson	2.4	3.1	-2.0	2.4	1.3	-0.1	-1.6	-0.6	-1.1	-2.0	-1.2	-0.7	-5.0
Frontier	0.1	-2.1	0.0	2.8	2.6	-0.4	-4.0	0.3	1.3	2.3	-1.8	-.26	-4.8
Hall	-6.1	-4.3	0.6	-1.6	-2.0	-0.8	0.2	0.4	3.8	2.3	1.7	5.4	+0.4
Hamilton	3.3	3.3	2.8	0.2	-3.2	-1.2	-3.2	-1.0	1.9	-1.6	-1.5	0.6	-4.3
Howard	-3.1	4.2	3.9	1.4	-1.7	0.7	1.9	1.7	-2.7	-2.0	-2.7	-1.6	17.1
Merrick	1.9	0.7	1.2	2.4	-2.9	-2.1	-0.4	-2.6	-0.4	-0.3	1.2	1.5	-3.3
Polk	2.2	-0.7	1.0	1.0	-3.2	-0.4	2.0	-2.4	-2.5	1.1	0.6	1.4	-0.6
Red Willow	3.2	-0.3	-0.5	-1.6	4.2	3.6	1.3	0.8	-1.1	-2.6	-1.6	-5.7	+1.6
...	-4.5	-1.5	-4.9	16.8	19.6	2.1	0.8	6.8	-10.6	-6.1	-3.9	14.0	37.0

table 8 of chapter 6

Numerical Residuals of Table 7 Arrayed by Adjustments, With Thirding Lines

(10 Counties by 12 Elections)

37.0	-10.6	-6.1	-4.9	-4.5	-3.9	-1.5	0.8	2.1	6.8	14.0	16.8	19.6	..County...
17.1	-2.7	-2.0	3.9	-3.1	-2.7	4.2	1.9	0.7	1.7	-1.6	1.4	-1.7	Howard
1.6	-1.1	-2.6	-0.5	3.2	-1.6	-0.3	1.3	3.6	0.8	-5.7	-1.6	4.2	Red Willow
0.4	3.8	2.3	0.6	-6.1	1.7	-4.3	0.2	-0.8	0.4	5.4	-1.6	-2.0	Hall
0.2	2.7	2.0	-3.4	-4.2	2.2	-2.9	1.2	0.1	1.6	2.4	-5.1	3.1	Adams
-0.6	-2.5	1.1	1.0	2.2	0.6	-0.7	2.0	-0.4	-2.4	1.4	1.0	-3.2	Polk
-1.3	-2.3	0.1	-3.3	0.8	2.1	-1.2	2.3	1.1	1.3	0.4	-1.6	1.3	Buffalo
-3.3	-0.4	-0.3	1.2	1.9	1.2	0.7	-0.4	-2.1	-2.6	1.5	2.4	-2.9	Merrick
-4.3	1.9	-1.6	2.8	3.3	-1.5	3.3	-3.2	-1.2	-1.0	0.6	0.2	-3.2	Hamilton
-4.8	1.3	2.3	0.0	0.1	-1.8	-2.1	-4.0	-0.4	0.3	-2.6	2.8	2.6	Frontier
-5.0	-1.1	-2.0	-2.0	2.4	-1.2	3.1	-1.6	-0.1	-0.6	-0.7	2.4	1.3	Dawson
Elect:	'52	'56	'28	'20	'60	'24	'44	'40	'48	'64	'32	'36	

No one can look at 54 = 9x6 more or less randomly arranged numbers and summarize them effectively. This is almost equally true of 27 = 9x3 numbers. Yet we want to look at the four corners of table 8.

Our real need is to compare the four corners. Each corner can be treated as if it were a batch. So we might as well use the schematic summaries of chapter 3 to contrast the SE and NW corners against the NE and SW corners. Table **9** gives the stem-and-leaf arrays. Plot **10** gives the schematic summaries. This plot shows little sign of structure, certainly little sign of tilt between the two pairs.

Since the numbers of residuals in the corners are so far from equal, it is natural to consider combining NW with SE and NE with SW. No obvious reason not to do this can be seen in plot 10, so plot **11** compares the pooled batches of residuals. The match is extremely close. We can only come to the conclusion that this set of data shows no trace of indication that a change of expression would help.

Another example

Now that we know that a diagonally-crossed pattern in the residuals can be taken as a signal of a need for change, we can usefully recall seeing this pattern for one of the sets of data with which we introduced coded residuals -- Daniels's thread counts. Table **12** shows the elementary analysis rearranged and thirded. This time neither rows nor columns have a middle portion. The stem-and-leaf arrays for the four corners seem to tilt upward a little from NW & SE to SW & NE. Plot **13** shows the schematic summaries, which give the same impression.

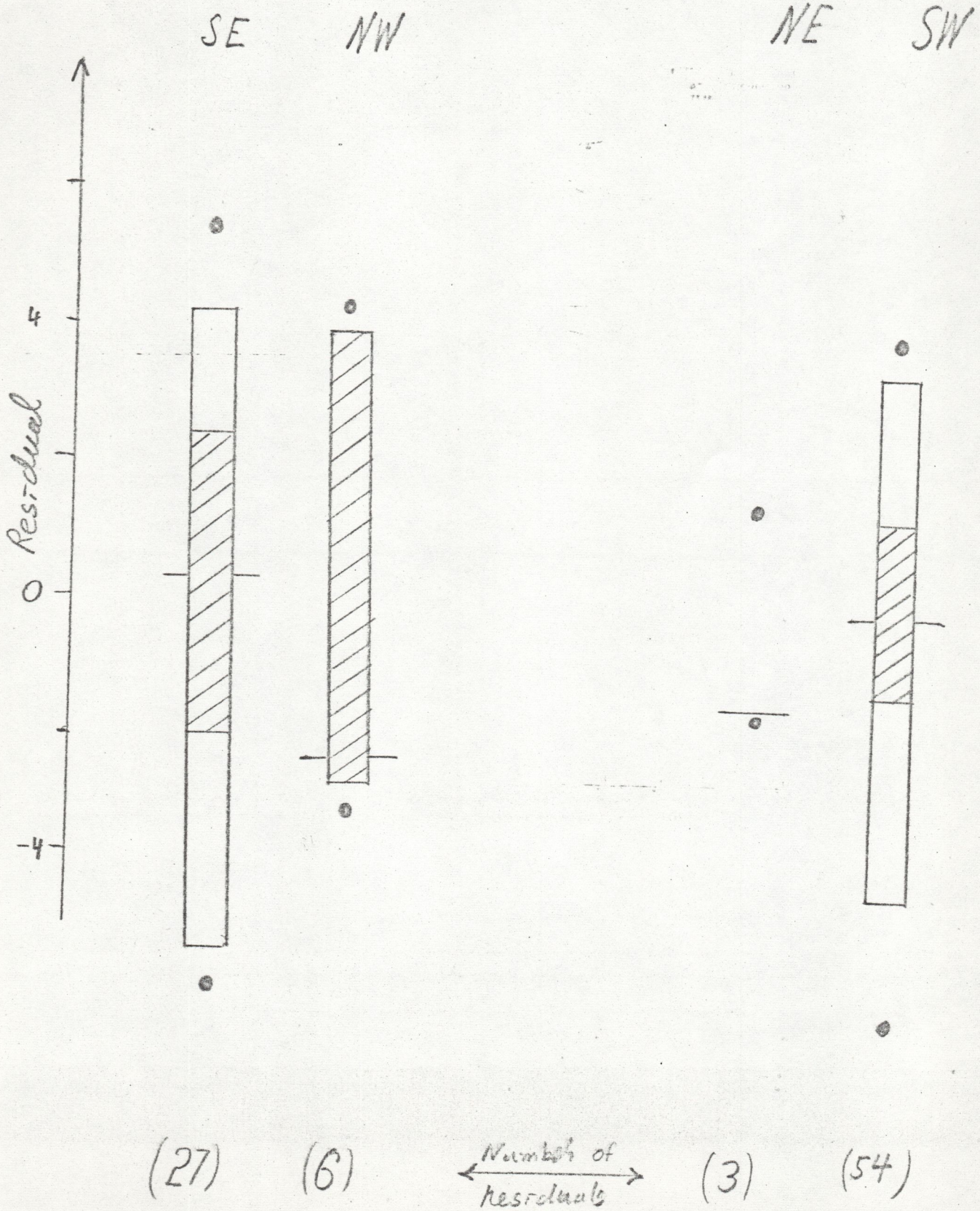
table 9 of chapter 6

Stem-and-Leaf Arrays for the Four Corners of Table 8
With Medians or Near-medians and Hinges Bold

	SE	NW	NE	SW
5	4			
4	2	2		
3	1	**9**		12338
2	44468			1223348
1	03345		**4**	01223799
0	246			0116678
-0	7			33457
-1	666		**67**	11225668
-2	069	**077**		000135679
-3	22	1		45
-4				23
-5	17			
-6				1
count:	27	6	3	54
Skeleton midmean:	+4	-9	-9	-1

plot **10** of chapter 6

Schematic summaries for the corners of table 8



plot 11 of chapter 6

Results of fusing opposite corners in table 8.

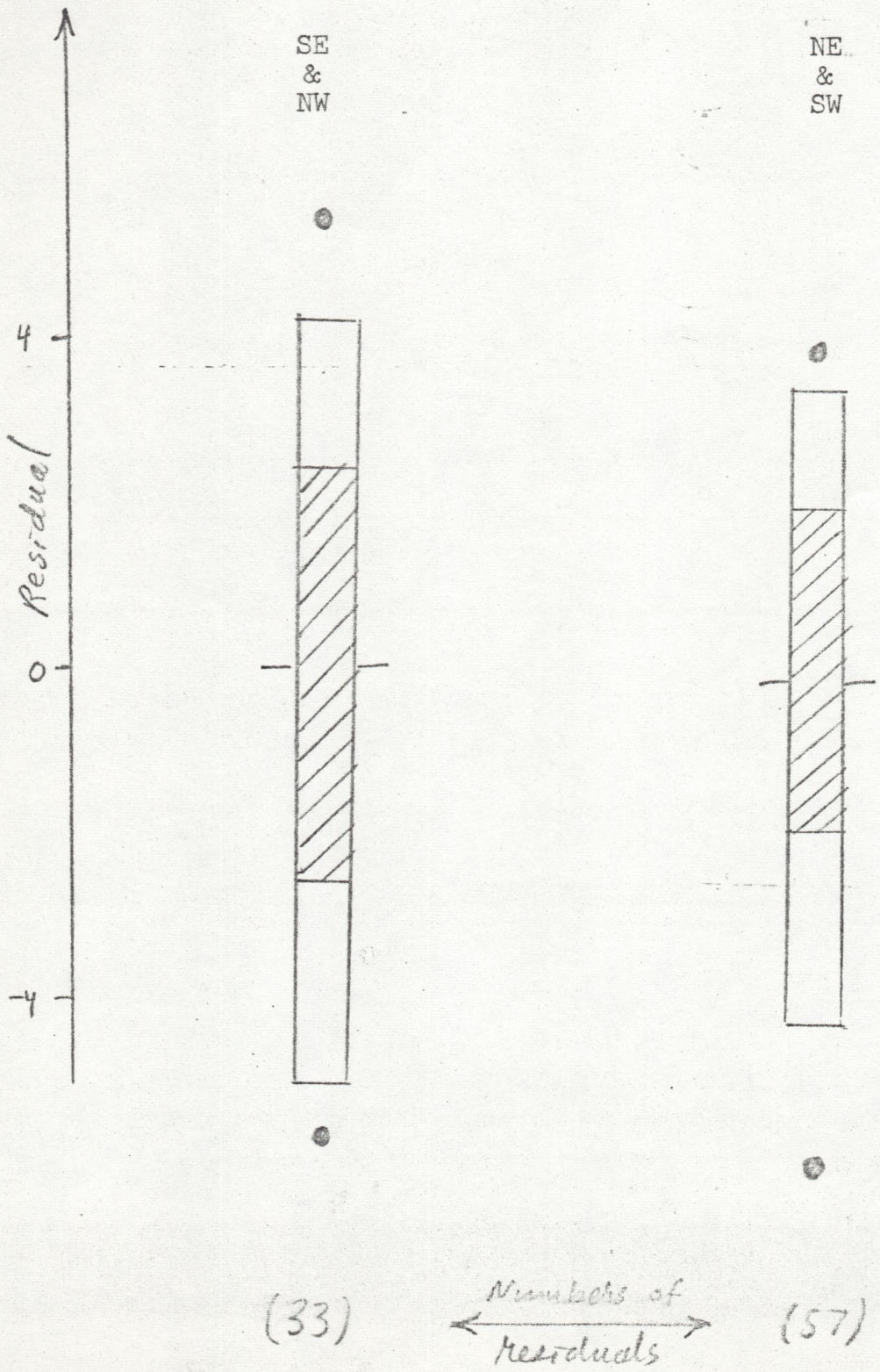


table 12 of chapter 6

Daniels's Thread Weights (See table 4 of Chapter 5)
and Their Rearranged Elementary Analysis

a) DATA, IN REARRANGED ORDER

				Ends*
40.42	41.42	42.18	42.28	C
40.22	41.47	42.05	41.95	D
40.27	41.28	41.63	41.19	B
40.27	41.18	40.22	40.99	E
38.43	39.08	38.83	39.35	A
Bobbin : 1	3	2	4	

b) ELEMENTARY ANALYSIS, REARRANGED AND THIRDED

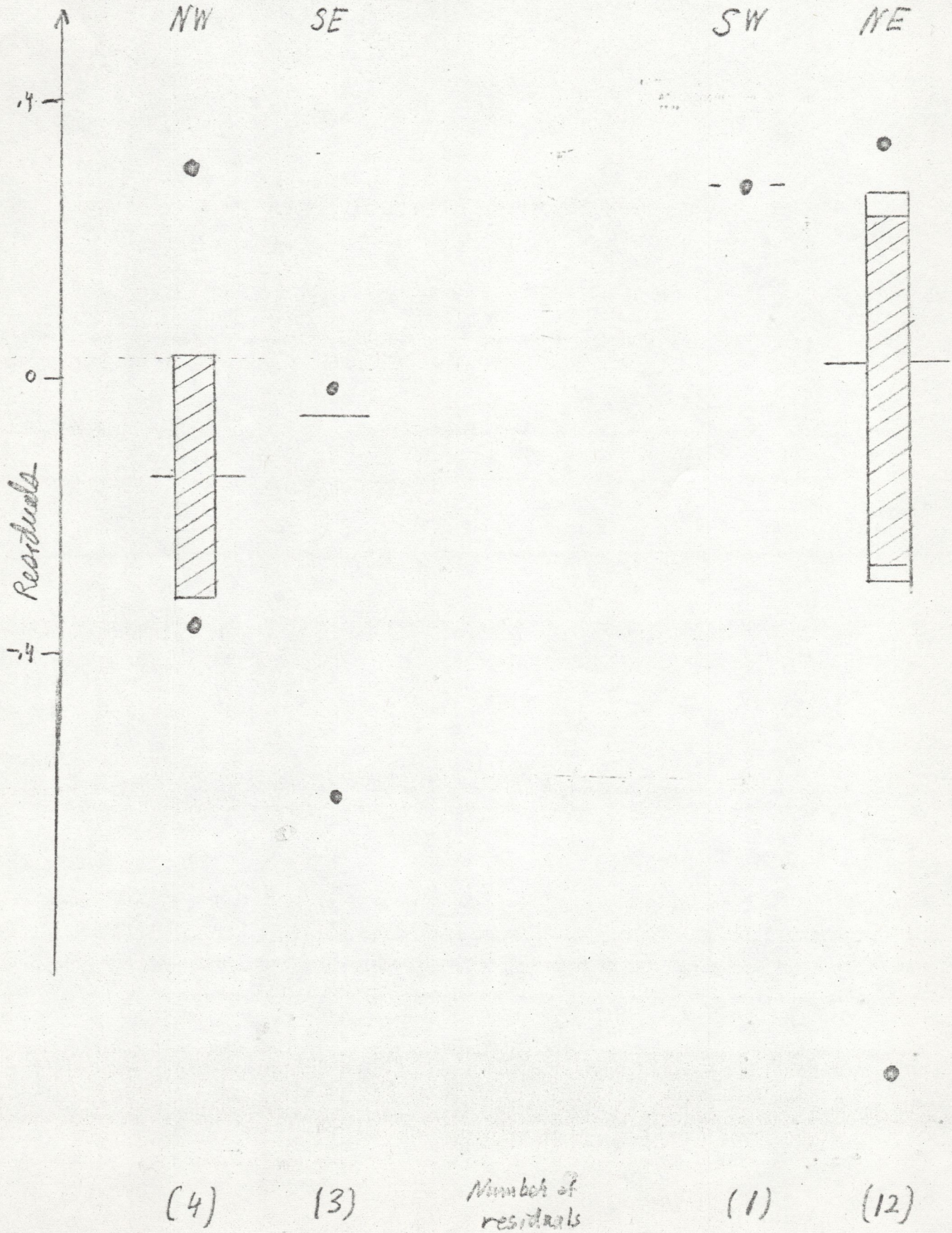
					Ends*
41.34	-1.08	-.10	.10	.13	
.48	-.32	-.30	.26	.33	C
.32	-.36	.09	-.29	.16	D
-.02	.03	.06	.21	-.26	B
-.29	.30	.23	-.93	-.19	E
-2.10	.27	-.06	-.51	-.02	A
Bobbin :	1	3	2	4	

c) STEM-AND-LEAF ARRAYS FOR THE FOUR CORNERS

	NW	SE	SW	NE
.4				
.3	0			3
.2			7	136
.1				6
.0	3			69
-.0		26		
-.1				9
-.2				69
-.3	26			0
-.4				
-.5		1		
-.6				
-.7				
-.8				
-.9				3
Midmean:	-.14	-.13	.27	.01

*in groups of 5.

Schematic summaries for the four corners of residuals in table 12(b)



What if this appearance were real? What if we tried to eliminate it with a change of expression for y? Since the ++ and -- corners are high, as in table 6(d), we should try lower powers than +1.

The row values vary over such a narrow range -- 38.43 to 42.28, less than ±5% from 40.35 -- that most of the usual reexpressions can have only rather small effects. Suppose we were to go so far as to try some convenient form of -1/y? If we make that form

$$z_{-1} = 82 - \frac{41^2}{y}$$

we obtain the following results:

y:	39	40	41	42	43
z_{-1}:	38.90	39.98	41	41.98	42.91

For $39 \leq y \leq 43$, the greatest change in going from y to z_{-1} is 0.10 (actually 0.105). This is unlikely to be enough to bring our four corners into line, but we shall try.

Table 14 shows the effects of the change. Panel (d) shows the change in the corner midmeans, and in the difference between the sums of these midmeans for diagonally opposite corners. This last is probably our simplest indicator of the apparent need for a change of expression. Going from y to -1/y lowered this from .55 to .36, leaving us about twice as far again to go. We changed power from +1 to -1, a step of 2, to go from y to -1/y, thus we might need a further step of 4, to a power of -5 to bring this indicator close to zero. The thought of using

table 14 of chapter 6

Effects of Reexpressing y as z_{-1}.
Compare and relate to table 12

a) CHANGES IN THE DATA, IN REARRANGED ORDER

					Ends*
	-.008	-.004	-.033	-.039	C
	-.015	-.005	-.026	-.021	D
	-.013	-.002	-.009	-.001	B
	-.013	-.001	-.015	-.000	E
	-.172	-.094	-.122	-.069	A
Bobbin	1	3	2	4	

b) ELEMENTARY ANALYSIS AFTER CHANGE, ASSUMING NO CHANGE IN PARTIAL DESCRIPTION

41.28	-1.09	-.10	.10	.13	Ends*
.48	-.328	-.304	-.227	.291	C
.32	-.375	.085	-.264	.139	D
-.02	-.017	.058	.201	-.261	B
-.29	.287	-.229	-.945	-.190	E
-2.10	.098	-.154	-.632	-.089	A
Bobbin	1	3	2	4	

c) STEM-AND-LEAF ARRAYS AFTER CHANGE

	NW	SE	SW	NE
.4				
.3				
.2	9			039
.1			0	4
.0				68
-.0	2	9		
-.1		5		9
-.2				366
-.3	38			
-.4				
-.5				
-.6		3		
-.7				
-.8				
-.9				4
Midmean**:	-.172	-.158	.098	-.066

*in groups of 5.
**based on three-decimal values.

table 14 (Cont'd) - 2

d) COMPARISON OF CORNER MIDMEANS

	NW	SE	SW	NE	(***)
for y	-.14	-.13	.27	.01	(.55)
for z_{-1}	-.172	-.158	.098	-.066	(.362)

*** = (SW+NE) - (NW+SE)

$$-\frac{1}{y^5}$$

or any similar form of expression is not a happy one. Given the limited amount of evidence we have, however, we should not let it bother us too much. This, like all our other results, is only an indication. We require some form of confirmation -- which may be quite informal -- before we have an obligation to believe and act upon it -- though we will often gain by following unconfirmed indications. (The strongest form of informal confirmation would be seeing a similar indication again and again in similar data gathered under differing circumstances.) Daniels gave several other sets of data in his paper. Clearly we need:

- a way to combine the evidence about change of expression from several sets of data.
- a way to compare, at least roughly, how well the sets tend to support one another.

To do something as complex as this easily and effectively, we are likely to do better with a graphical procedure.

6D. Looking at the evidence

It is far from crystal clear how to look at the residuals to see what they seem to say about the need for a change in expression. The only convincing way to see what to do involves a fair amount of detailed argument. To put such a treatment here would be too much of an interruption, yet some readers may want to see it. Accordingly, we have stowed it out of the way in sections 6R, 6S and 6T.

The main results are simple:

- We can do a good job of looking at the evidence about the need for change in expression, at least for a change to another power or to logs, by plotting the residuals against suitable comparison values.

- If we have one set of data, we may take

(row adjustment)x(column adjustment)

as our comparison value.

- If we have several sets of data, each of whose values hover around nearly the same value, we can do the same.

- If we have several sets of data hovering around different values, we can use

$$\frac{\text{(row adjustment)x(column adjustment)}}{\text{(general typical for that set of data)}}$$

as our comparison value.

- The stronger any apparent dependence of residual on comparison value, (i) the stronger the evidence that a change in expression is needed, and (ii) the larger reduction in size of residuals to be expected to accompany the removal of apparent structure when a better choice of expression is used.

Table 15 applies these precepts to the same thread weights reported by Daniels that we have seen so many times. We have kept only a minimum number of figures in order to speed (and make safer) our arithmetic. (Recall that using thirds corresponds to using only low, center, and high as values.) Plot 16

table 15 of chapter 6

Comparison Values for Residuals in Daniels's Thread Weights

(See table 12 above)

a) COMPARISON VALUES FOR USE NEAR 40

Ends		Bobbin 1	2	3	4
		-1.0	0.1	-0.1	0.2
A	-2.1	2.10	-0.21	0.21	-0.42
B	0.0	0.00	0.00	0.00	0.0
C	0.4	-0.40	0.04	-0.04	0.8
D	0.3	-0.30	0.03	-0.03	0.6
E	-0.3	0.30	-0.03	0.03	-0.6
	...				

Note. Bordering values are adjustments from elementary analysis. Cp. (c) below.

b) 100 TIMES COMPARISON VALUES FOR ANY USE

A	5.2	-.5	.5	-1.0
B	0	0	0	0
C	-1.0	.1	-.1	.2
D	0.8	.1	-.1	.2
E	-0.8	-.1	.1	-.2

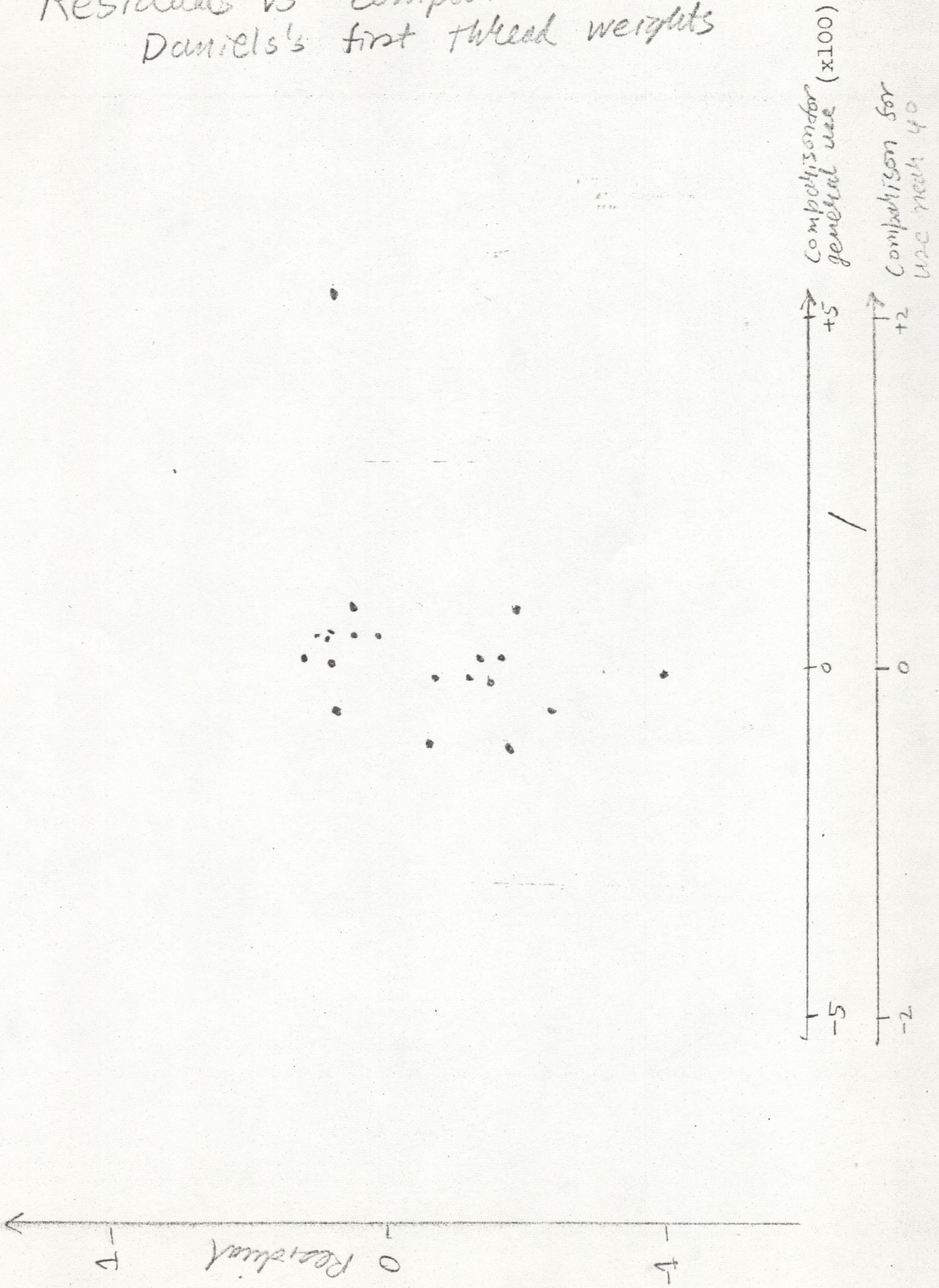
Note. Values are 100/40 times values in (a).

c) RESIDUALS, BY MIDMEAN DIFFERENCES

A	.17	-.59	-.14	-.13	-2.11
B	-.07	.14	-.02	-.37	-.03
C	-.40	.21	-.36	.24	.45
D	-.47	.22	-.18	.04	.32
E	.19	-1.00	.14	-.31	-.29
	-1.05	0.09	-0.09	0.17	41.42

plot 16 of chapter 6

Residuals vs comparison values
Daniels's first thread weights



shows how the residuals look when plotted against either set of comparison values. The two horizontal scales show clearly the equivalence of the two comparison values -- so long as we deal with one set of data or several sets of data with about the same general typical.

Looking at plot 16 we see that most of the evidence about the need for a better expression comes from the one value on the far right, which suggests -- but does not urge -- a tilt upward to the right. Were we to put this value out of our minds, the values that remain also suggest, much more faintly, a tilt of the same sign (high to right). Now that we have our evidence in the form of a simple plot we can do a much better job of gaining a feeling about the strength of its suggestion and the number of residuals that contribute significant evidence.

In addition to the set of data we have been analyzing, Daniels also reported a second set of data of the same pattern for which the carding was done 65% faster.

Table 17 sets out the data, elementary analysis, and comparison values for Daniels's second set of data. Plot 18 combines both sets, showing the residuals against the comparison values. The hint of upward slope is now a little weaker.

Daniels's paper also contained the results of four further tests carried out with a very much heavier thread, with weights of five ends near 104 grams rather than near 41 grams. Table 19 collects the residuals according to comparison value and calculates medians where appropriate. Plot 20 summarizes all six sets of residuals.

table 17 of chapter 6

Daniels's Second Set of Thread Weights

a) DATA, WEIGHTS IN GRAMS FOR GROUPS

End group	Bobbin 1	2	3	4
A	38.28	38.90	38.85	39.50
B	39.15	40.36	39.51	39.88
C	39.03	40.53	39.74	41.06
D	39.37	41.11	40.63	41.41
E	40.00	39.76	40.41	41.02

b) ELEMENTARY ANALYSIS OF (a)

					(*)
A	.21	-.20	.16	.03	-1.20
B	.26	.43	-.01	-.42	-0.37
C	-.23	.22	-.16	.38	0.01
D	-.47	.22	.15	.15	0.59
E	.38	.90	.16	-.01	0.36
(*)	-.83	.21	-.20	.58	40.09

c) REARRANGED ANALYSIS

40.09	-.83	-.20	.21	.58	Ends
0.59	-.47	.15	.22	.15	D
0.36	.38	.16	.90	-.01	E
0.01	-.23	-.16	.22	.38	C
-0.37	.26	-.01	.43	-.42	B
-1.19	.22	.16	-.20	.03	A
Bobbin	1	3	2	4	

d) COMPARISON VALUES FOR USE NEAR 40

				Ends
-.48	-.12	.12	.36	D
-.24	-.08	.08	.24	E
0	0	0	0	C
.32	.08	-.08	-.24	B
.96	.24	-.24	-.72	A
Bobbin: 1	3	2	4	

table 17 (Cont'd) - 2

e) 100 TIMES COMPARISON VALUES FOR GENERAL USE

					Ends
	-1.2	-.3	.3	.9	D
	-0.6	-.2	.2	.6	E
	0	0	0	0	C
	.8	.2	-.2	-.6	B
	2.4	.6	-.6	-1.8	A
Bobbin:	1	3	2	4	

Source: H. E. Daniels 1938. Some problems of statistical interest in wool research. Supplement to the Journal of the Royal Statistical Society 5: 89-112, specifically Test II of Table A on page 109.

plot 18 of chapter 6

Residuals vs comparison values
Daniels's two sets of thread weights
combined

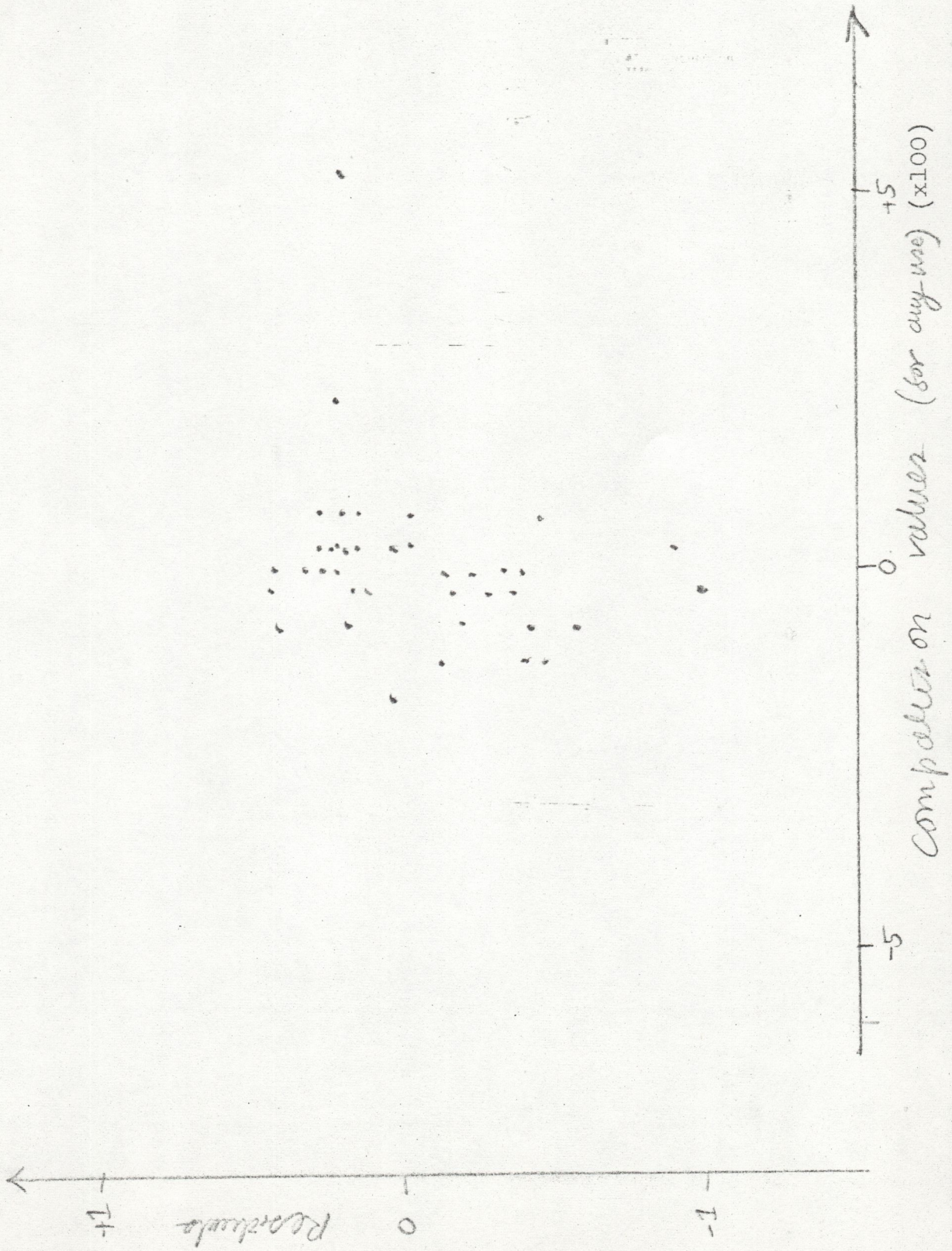
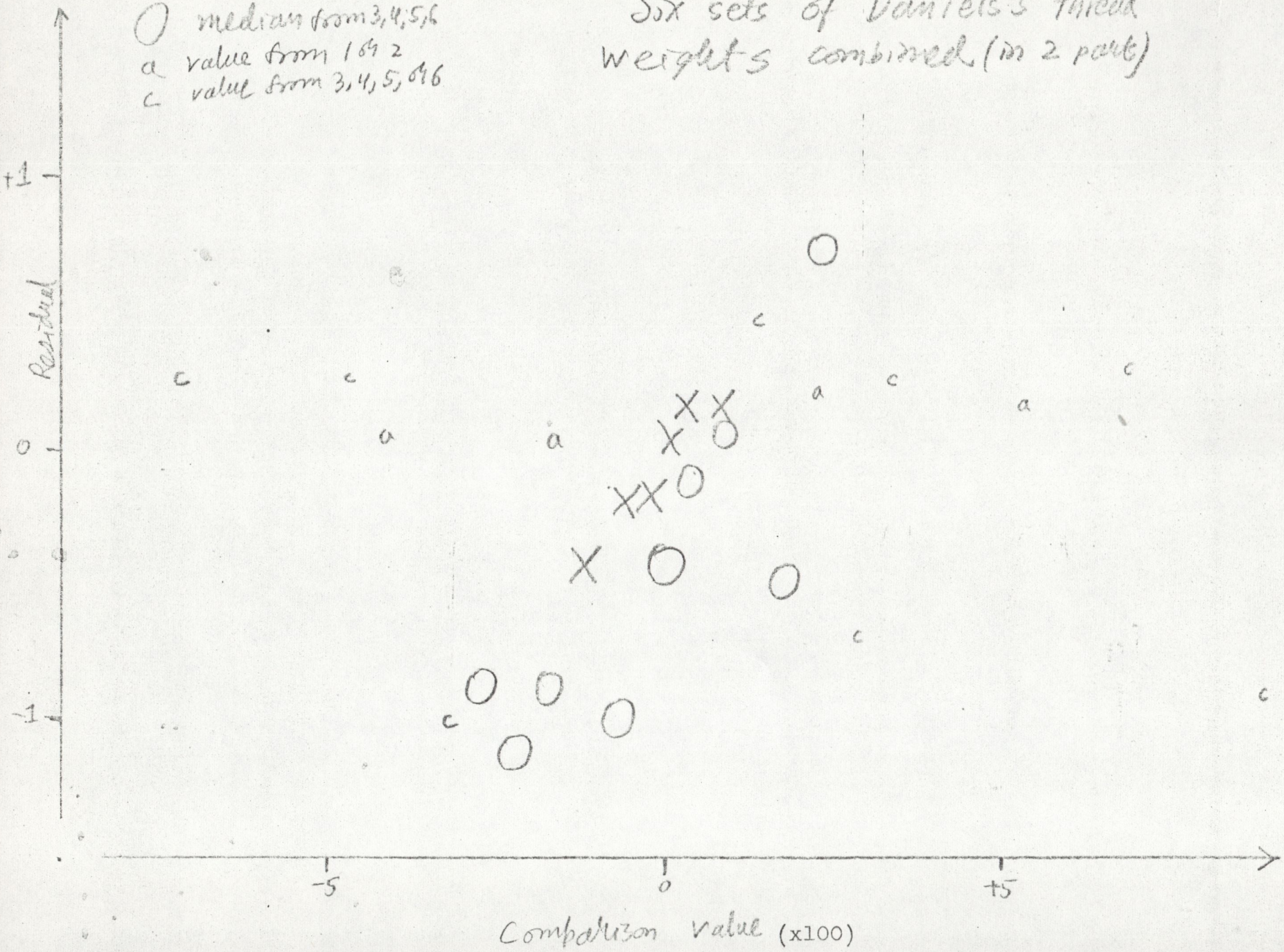


table 19 of chapter 6

Residuals for Four More Sets of Thread Weight Data
Arranged by Comparison Value

Comparison	Residuals	Median
+9.5 to +9.9	-.37	
+8.5 to +8.9	-.91	
+6.5 to +6.9	.30	
+3.0 to +3.4	.24	
+2.5 to +2.9	-.71	
+2.0 to +2.4	-.86,+.55,+.70,+.87,+.93,+1.09	.78
+1.5 to +1.9	-1.98,-.49,-.22	-.49
+1.0 to +1.4	.48	
+0.5 to +0.9	-1.70,-1.67,-1.34,-.87,-.20,.04,1.07, 1.07,1.12,1.28,1.32	.04
+0.1 to +0.4	-1.14,-1.01,-.69,-.66,-.62,-.16,-.08,.04, .14,.23,.51,.60	-.12
0.0	-.95,-.47,-.44,-.24	-.46
-0.1 to -0.4	-1.93,-1.70,-1.01,-.67,-.57,-.27,-.08, -.01,.13,.15,.15,.54	-.18
-0.5 to -0.9	-2.21,-1.52,-1.52,-1.06,-1.02,-.99,-.39, .53,.65,.67	-1.00
-1.0 to -1.4	none	
-1.5 to -1.9	-1.38,-.90,-.88,-.55,-.32	-0.88
-2.0 to -2.4	-1.40,-1.21,.99	
-2.5 to -2.9	-.82,-.74,.09	
-3.0 to -3.4	-.99	
-4.0 to -4.4	-.35,.26	
-4.5 to -4.9	.28	
-7.0 to -7.4	.26	

plot 20 of chapter 6



The large characters in plot 20, representing medians of at least three values, tend to show increasing residual with increasing comparison value. The single values at low and high comparison values do nothing of the sort. If we had three or four times as much data, we could probably tell how the residuals are behaving. As it is, we wonder.

As this example suggests, firm evidence about choice of expression sometimes requires a substantial amount of data. This is especially true when, as in this thread-weight example, substantial changes in the form of expression make relatively small changes in the residuals. It is, however, not always true.

As table 3 so clearly showed for the Winsor and Clarke data, when we have good "leverage" a single 3x10 table may offer extremely strong evidence that a change in form of expression is needed. The amount of "leverage" offered by data is not fixed by the size and shape of the table or tables that we have at hand.

Behavior that can be simplified by an extreme change of expression may sometimes be better handled by a change in form of partial description. The slope corresponding to the medians shown in plot 20 is about 30. This would correspond to either:

- describing $-1/(\text{weight})^{29}$ in terms of row adjustment plus column adjustment , or
- describing weight in terms of

$$(\text{row adj}) + (\text{col adj}) - 30\,\frac{(\text{row adj})(\text{col adj})}{(\text{general typical})}$$

The second of these choices, though formally more complex, might be preferable in this instance.

6E. Stripes and skews

For two-way comparisons -- and for one-way comparisons in blocks -- the sort of evidence for a change of expression we have been considering is of overriding importance. In this it is like making simple plots first straight and then flat. If it offers an adequate basis for choice of expression, we should follow its guidance.

The way that comparison values are formed shows what is needed for there to be a solid basis. We need enough difference among comparison values to give a tendency to tilt a fair chance to show itself. Since comparison values are products of the form "row adjustment times column adjustment" this means that both factors -- both row and column adjustments -- should vary enough to let their products vary enough to give us a good grasp on the situation. If either all row adjustments or all column adjustments are nearly alike, we are going to find it hard to be guided to a good choice of expression.

If the first line of guidance will not work, we must fall back to the second line. Here, as in the comparison of several batches, the second line of guidance involves looking at spreads and seeking a form of expression that tends to make spreads more equal.

Equal among what? Equal among stripes, whether these be rows or columns. Spread of what? Spread of the residuals within a stripe.

If one kind of stripe has widely spread adjustments, a different choice of expression will lead to a different

pattern of spreads for these same stripes. This will be worth our attention if the other kind of stripes has very similar adjustments. This tigerish appearance, with stripes one way strong and the other almost invisible, can and does occur with two-way comparisons. Moreover, it occurs so frequently as to deserve the word "rampant" when we deal with one-way comparisons in blocks.

Let us turn to data of Geoffrey Beale on numbers of insects, here Leptinotarsa decemlineata, on plots given 4 different treatments. Here treatments are strong, while blocks are quite weak. If we neglect blocks, and act as if we had only one classification, we need only calculate a measure of level and a measure of spread for each treatment. Table 21 does this, using means as measures of level and midspreads as measures of spread. Increasing level is clearly accompanied by increasing spread. This leaves us unhappy with raw counts, and ready to try lower powers. Tables 22 and 23 repeat the calculation for:

- the square roots of counts.
- the logs of counts.

Comparing these three results, we clearly take the square root if we must take one of the three, since its apparent dependence of midspread on mean is by far the weakest.

We might look at the same data as a two-way table, treating block-to-block differences as something that must be adjusted if we are to learn about forms of expression. We

table 21 of chapter 6

Treatment Midspreads of Beale's Counts of Insects, *Leptinotarsa decemlineata*, for Pairs of Plots Combined and Blocks Disregarded

	Treatment			
	1	3	2	4
a) DATA Block				
1	512	423	190	19
2	755	295	100	44
3	1059	680	46	56
4	882	681	49	12
5	580	182	7	2
6	423	343	22	13
7	851	530	11	7
b) SUMS, MEANS, HINGES AND MIDSPREADS				
Sum	5062	3134	425	153
Mean	723	448	61	22
U hinge	882	680	100	44
L hinge	512	295	11	7
Midspread	370	385	89	37

Source: Geoffrey Beale 1942. The transformation of data from entomological field experiments so that the analysis of variance becomes applicable. Biometrika 32: 243-262, specifically table 3 on page 245.

table 22 of chapter 6

Treatment Midspreads for Square Roots of Beale's Insect Counts with Blocks Disregarded

(Data entries are 10 $\sqrt{\text{count for two plots}}$)

a) DATA

Block	Treatment 1	Treatment 3	Treatment 2	Treatment 4
1	224	204	136	44
2	276	172	100	66
3	330	264	68	74
4	300	258	70	35
5	240	136	26	14
6	204	184	47	36
7	294	232	33	26

b) SUMS, MEANS, HINGES AND MIDSPREADS

	Treatment 1	Treatment 3	Treatment 2	Treatment 4
Sum	1868	1452	480	295
Mean	267	207	69	42
U hinge	300	258	100	66
L hinge	224	172	33	26
Midspread	76	86	67	40

table 23 of chapter 6

Treatment Midspreads for Logs of Beale's Insect Counts
With Blocks Disregarded
(Data entries are 100 $\log_{10}$(count))

	Treatment			
Block	1	3	2	4
a) DATA				
1	271	262	228	128
2	288	247	200	164
3	303	284	166	175
4	295	283	169	108
5	276	226	85	30
6	263	254	134	111
7	293	272	104	85
b) SUMS, MEANS, HINGES AND MIDSPREADS				
Sum	1989	1828	987	801
Mean	284	261	141	114
U hinge	295	283	200	164
L hinge	271	247	104	85
Midspread	24	36	96	79

could do a complete elementary analysis to find our residuals, but a shortcut is simpler.

When we want only the midspreads of residuals by stripe, the streamlined calculation introduced toward the end of chapter 5 allows us to obtain the desired result quickly, provided analysis by means followed by midmean polish is good enough for us. Indeed, only one-half of the streamlined calculation is needed.

If we want midspreads by column, we need only:

- find and subtract column means.
- find and subtract row midmeans of the result.
- find midspreads of the second result.

We can do this, since further adjustment by columns will not change the column midspreads.

Table 24 goes through the calculation for the counts. Again increasing level seems to mean increasing spread. Table 25 summarizes the means and midspreads for suitable versions of count, $\sqrt{\text{count}}$, and log count, adding fourth and eighth roots (easily found by successive square rootings) in an attempt to tie down matters further.

The ordering of the treatments as to level is clear, but the irregularities among the midspreads is great enough to make us think that more than four midspreads should be examined to settle such a question.

If we have to make some choice, using square roots still seems plausible.

table 24 of chapter 6

Streamlined (for Hand Calculation) Calculation of Spreads of Residuals for Beale's Leptinotarsa decemlineata Counts, Blocks Considered

(Data Entries Are Counts)

	Treatment			
Block	**1**	**3**	**2**	**4**
a) DATA				
1	512	423	190	101
2	755	295	100	44
3	1059	680	46	56
4	882	681	49	12
5	580	182	7	2
6	423	343	22	13
7	851	530	11	7
Sum	5062	3134	425	153
Mean	723	448	61	22

b) TREATMENT-ADJUSTED VALUES -- (adjusted for means)

					(*)
1	-211	-25	129	-3	-14
2	32	-153	39	22	27
3	336	232	-15	34	133
4	159	233	-12	-10	74
5	-143	-266	-54	-20	-98
6	-300	-105	-39	-9	-72
7	128	82	-50	-15	34

(*) = Midmean

c) DOUBLY-ADJUSTED VALUES -- further adjusted for midmeans of (b)

1	-197	-11	143	11
2	5	-180	12	-5
3	203	299	-148	-99
4	85	159	-86	-84
5	-45	-168	44	78
6	-228	-33	33	63
7	94	48	-84	-49

d) HINGES AND MIDSPREADS

U hinge	94	96	44	63
L hinge	-197	-168	-84	-84
	291	264	128	147

table 25 of chapter 6

Comparison of Midspreads for Various Forms of Expression for Beale's Data, Blocks Considered

	Treatment			
	1	3	2	4
a) MEANS				
y	723	448	61	22
$\sqrt{y}$	267	207	69	42
$\sqrt[4]{y}$	163	115	80	63
$\sqrt[8]{y}$	127	120	89	79
$\log y$	284	261	141	114
b) MIDSPREADS				
y	291	264	128	147
$\sqrt{y}$	60	55	56	28
$\sqrt[4]{y}$	32	26	20	13
$\sqrt[8]{y}$	12	6	12	10
$\log y$	25	12	42	61

It is almost -- but not quite -- true that we have looked for -- and tried to eliminate -- striped spread in exactly the same way that we tried to eliminate spread differences among batches. Only one relevant difference needs sign-posting:

- When dealing with batches we had a means of using the relation of spread to level to tell us what change in expression might be good to try.

The simple scheme that worked there will not work in a two-way table, since spread leaks from one stripe to another in the process of finding residuals. Table 26 shows a simple artificial example. In this extreme instance, all the spread was in the first two columns of data. Beginning there, it leaked, until the columns of residuals all show equal spread.

How much leakage will there be? This question is not easy to answer for analysis by midmean differences. For analysis by means, however, careful argument suggests replacing the spread observed for a stripe of residuals by

$$\sqrt{(\text{that spread})^2 - \frac{1}{s-1}\,(\text{sum of all } (\text{spread}^2)s)}$$

when there are s stripes and s is at least 3, and then treating these new values as if they were spreads of separate batches.

What we should do with analysis by midmean differences is not so clear. We could always avoid the issue by planning to repeat the analysis for enough different choices of expression to clear up the situation. This course is safe, though possibly laborious.

table 26 of chapter 6

Artificial Example of Spread Leaking From Column to Column

a) DATA

Rows	Columns 1	2	3	4
1	-12	-12	0	0
2	-8	-8	0	0
3	-4	-4	0	0
4	-4	-4	0	0
5	0	0	0	0
6	0	0	0	0
7	0	0	0	0
8	0	0	0	0
9	4	4	0	0
10	4	4	0	0
11	8	8	0	0
12	12	12	0	0

b) ELEMENTARY ANALYSIS

1	-6	-6	6	6	-6
2	-4	-4	4	4	-4
3	-2	-2	2	2	-2
4	-2	-2	2	2	-2
5	0	0	0	0	0
6	0	0	0	0	0
7	0	0	0	0	0
8	0	0	0	0	0
9	2	2	-2	-2	2
10	2	2	-2	-2	2
11	4	4	-4	-4	4
12	6	6	-6	-6	6
	0	0	0	0	0

c) COLUMN MIDSPREADS

U. hinge:	2	2	2	2
L. hinge:	-2	-2	-2	-2
Midspread:	4	4	4	4

Those who wish to live a little dangerously -- only a little, because by the time we try two or three choices of expression we are going to be led to the apparent answer, almost willy-nilly -- can try making half as much adjustment as for means, using

$$\sqrt{(\text{that spread})^2 - \frac{1}{2(s-1)}\,(\text{sum of all } (\text{spread})^2)}$$

as if these were spreads of batches. (Anything of this sort can only work as a quite weak compromise.)

Before we close, we should come back to Geoffrey Beale and his Leptinotarsa decemlineata. The data was stated to be for pairs of plots combined. We found it sensible to analyze the square roots of our counts. If this is true between plot-pairs, it is also likely to be true within plot-pairs. Accordingly, table 27 gives the data in more detail and analyzes the square roots of the counts. We see that while the midspreads do vary (as we should expect) the variation does not seem to be strongly attached to level. The square root still appears to be a reasonable choice.

It is conceivable that we may be pushed back to the third line of defense: choosing that expression that makes stripes of residuals look symmetrical. If this should come to pass, then we would be able to treat each set of stripes of residuals as if it were a set of batches. (Since there are no refined techniques for batches, there is nothing to change when we go to stripes.) Doing this is rarely needed, infrequently useful.

table 27 of chapter 6

The Leptinotarsa Decemlineata Counts in More Detail
(Two Counts Per Treatment Per Block)

Block	Treatment 1	2	3	4
a) DATA				
1	305, 207	97, 93	270, 153	7, 12
2	391, 369	49, 51	105, 190	42, 2
3	420, 639	21, 25	341, 348	34, 22
4	355, 527	12, 37	469, 212	8, 4
5	287, 293	3, 4	82, 100	1, 1
6	175, 248	10, 12	57, 285	10, 3
7	454, 397	10, 1	221, 309	4, 3
b) SQUARE ROOTS OF COUNTS (multiplied by 10; found from table 25 of chapter 2)				
1	176, 144	98, 96	164, 124	26, 35
2	196, 192	90, 72	102, 136	64, 14
3	204, 252	46, 50	184, 188	58, 47
4	188, 232	37, 60	216, 144	28, 20
5	168, 172	17, 20	90, 100	10, 10
6	132, 156	32, 35	76, 168	32, 17
7	212, 200	32, 10	148, 176	20, 17
c) SUMS OF SQUARE ROOTS (still multiplied by 10)				
1	320	194	288	61
2	388	172	238	78
3	456	96	372	105
4	420	97	360	48
5	340	37	190	20
6	288	67	244	49
7	412	42	324	37
Sum	2624	705	2016	398
Mean	375	101	288	57

d) ADJUSTED FOR COLUMN MEANS					(*)
1	-55	93	0	4	2
2	13	71	-50	21	17
3	81	-5	84	48	64
4	45	-4	72	-9	20
5	-35	-64	-98	-37	-50
6	-87	-34	-44	-9	-39
7	37	-59	36	-20	8

table 27 (Cont'd) - 2

e) ALSO ADJUSTED FOR ROW MIDMEANS (from (*))

1	-57	91	-2	2
2	-4	54	-67	4
3	17	-69	20	-16
4	25	-24	52	-29
5	15	-14	-48	13
6	-48	5	-5	30
7	29	-67	28	-28

f) HINGES AND MIDSPREAD

U hinge	25	54	28	13
L hinge	-48	-67	-48	-28
Midspread	73	121	76	41
(Means)	(375)	(101)	(288)	(57)

6F. How far have we come?

The most direct and convincing type of evidence from a two-way table that a change of expression is called for comes from "corner behavior". Such evidence -- or the absence of such evidence -- is clear when neither row adjustments nor column adjustments are nearly constant. As the Winsor and Clarke example shows, a single two-way table of moderate size can shout out its need for a change of expression. In such circumstances, it may well be only after such a change that we can see the other things that appear to be going on.

When our "leverage" is poor, so that we must listen for a strong whisper not a shout, careful inquiry into whether, and what, change of expression is needed uses an approximation for the effects of a change in form of expression. If this change is to the pth power of the original data, the residuals will be changed, approximately, by

$$(p-1)\ \frac{(\text{row adjustment})(\text{column adjustment})}{(\text{general typical})}$$

Thus plotting residuals against the result of suppressing the (p-1) factor offers a convenient way of looking to see:

- whether a change would help.
- about what size help is indicated.

As the Daniels example showed, such plotting:

- may involve several tables of data, which may have been gathered quite separately.
- may, or may not, require large amounts of data to reach clearly defined results.

In those instances which are weaker than a shout but stronger than a whisper, a procedure involving division of both

rows and columns into thirds lets us look at the residuals in the four corners of the table. Since the product of row adjustment and column adjustment will tend to be > 0 in one pair of opposite corners and < 0 in the other pair, the indication of need for change of expression is a tendency of residuals to be > 0 in one pair of opposite corners and < 0 in the other pair.

Only if the sort of analysis just described is inconclusive, is it appropriate to look for other cues which might call for a change of expression. Only the "corners of the rearranged table" of which the careful form is "plotting against comparison values" has an importance comparable with making plots first straight and then flat.

The second line of defense, to be used when the comparison values are too small to give good leverage, is to look at the midspreads of stripes of residuals and see if they seem to shift systematically with the level (e.g. the midmean, the median, or the mean) of their stripe. This analysis is only appropriate when the table has one set of strong stripes and one set of weak ones. This situation often arises when one striping corresponds to blocks. If the weak striping is weak enough, we can forget it entirely, treating each of the strong stripes as a separate batch of numbers. If it is not quite that weak, we can still find midspreads -- and, to adequate accuracy, levels -- without calculating the full elementary analysis and the corresponding residuals. One-half of the streamlined analysis set up in chapter 5 -- the one involving first analysis by means and then polish by midmeans -- does the trick.

Conceivably, if both stripes are weak, we might appeal to symmetry as a basis for choosing expression. This will, however, be quite infrequently helpful.

*****

6R. Why do we look at the evidence as we do?

What we want of a change in expression is that:

- the residuals cease to show an obvious need for a further change in expression.

- the residuals become "smaller" (where the quotation marks call for care in defining "smaller").

Since all our fitting procedures are transparent, in the sense of chapter 5, the residuals after the change can be equally well calculated in a variety of ways, including:

- residuals of z, the new values.

- residuals of (old residuals + changes, z-y)

- residuals of (old residuals + residuals of changes), which will often be close to (old residuals) + (residuals of changes).

Thus any comparison may usually be satisfactorily focused on what is simplest to compute. This involves comparing:

1) old residuals, with

2) residuals of changes, or something in approximately fixed proportion thereto.

If we can approximate a constant multiple -- or constant fraction -- of the changes, and then find the residuals corresponding to the table of approximate changes, we can -- and we should -- compare the old residuals with the values thus obtained.

Just as we picked

$$y - \frac{(y-c)^2}{y}$$

as the linear coding of $-1/y$ that is "close" to y for y near c, so it turns out that there is a linear coding of any power, y^p that is close to y for y near c. In section 6S we look into the existence and properties of this new expression, finding that, just as for

$$y - \frac{(y-c)^2}{y}$$

which we can approximate by

$$y - \frac{(y-c)^2}{c}$$

that the following hold:

- given one set of data, the effect of going from y to the coded form of y^p is, for any p, a change roughly proportional to $(y-c)^2$.
- given several sets of data for which the same value of c is appropriate, the same is true.
- given several sets of data for which different values of c are appropriate, the effect of going from y to the coded form of y^2 is, for any fixed p, a change roughly proportional, with coefficient p-1, to $(y-c)^2/c$.

If we are to use the simplest of these approximate results, we should find:

- the residuals of $(y-c)^2$ and compare the original residuals with this result.

If y is to be the observed value, then

y = (general typical) + (row adjustment) + (column adjustment) + (residual)

and it is clearly sensible to take c = (general typical). Thus, expanding and grouping,

$$(y-c)^2 = (\text{row adjustment})^2 + (\text{column adjustment})^2$$
$$+ 2(\text{row adjustment})(\text{column adjustment})$$
$$+ 2(\text{row adj})(\text{residual}) + 2(\text{col adj})(\text{residual})$$
$$+ (\text{residual})^2$$

The two terms in the first row can be exactly fitted. Since our fitting procedures are transparent, they will disappear from the residuals of the changes. Thus we can forget them.

The last three terms each involve the residuals as a factor or factors. We can hope that they are small. They often are.

If we had started with the changes corresponding to fitted, rather than observed values, the term "+ residual" would have disappeared from y and these three terms would never have been there. This comment may help to explain why these terms are often small.

This leaves the middle line, from which we can cross out the factor of 2 without loss, leaving:

$$(\text{row adjustment}) \times (\text{column adjustment})$$

What will happen when we find the residuals of the table described by this expression? For reasons given

in section 6T, this table will turn out to be its own residual.

The natural successions of these steps of approximation are shown in picture 28. Following either route, the end is the same.

Table 29 shows numerically, for our artificial example, how these processes can behave. In this instance, using

(row adjustment)x(column adjustment)

works well.

Thus we have learned that:

- given one set of data, a natural way to look at its residuals for an indication of a need for a change in expression is to compare them with the products of row adjustment and column adjustment.

- given several sets of data, all of which vary around nearly the same value; the same result holds.

- given several sets of data, some varying around lower values, some around higher ones, it is natural to compare residuals with

$$\frac{\text{(row adjustment)(column adjustment)}}{\text{(general typical)}}$$

We have thus reached the answers we sought.

6S. Matched powers, how do they differ?

When we are concerned with several forms of expression, for values which lie moderately close to some numerical value, say c, it is often very convenient to work with

picture 28 of chapter 6

The Two Stages of Approximation Leading From "Residuals of Actual Changes in Observed Values" to "Product of Row Adjustments and Column Adjustments

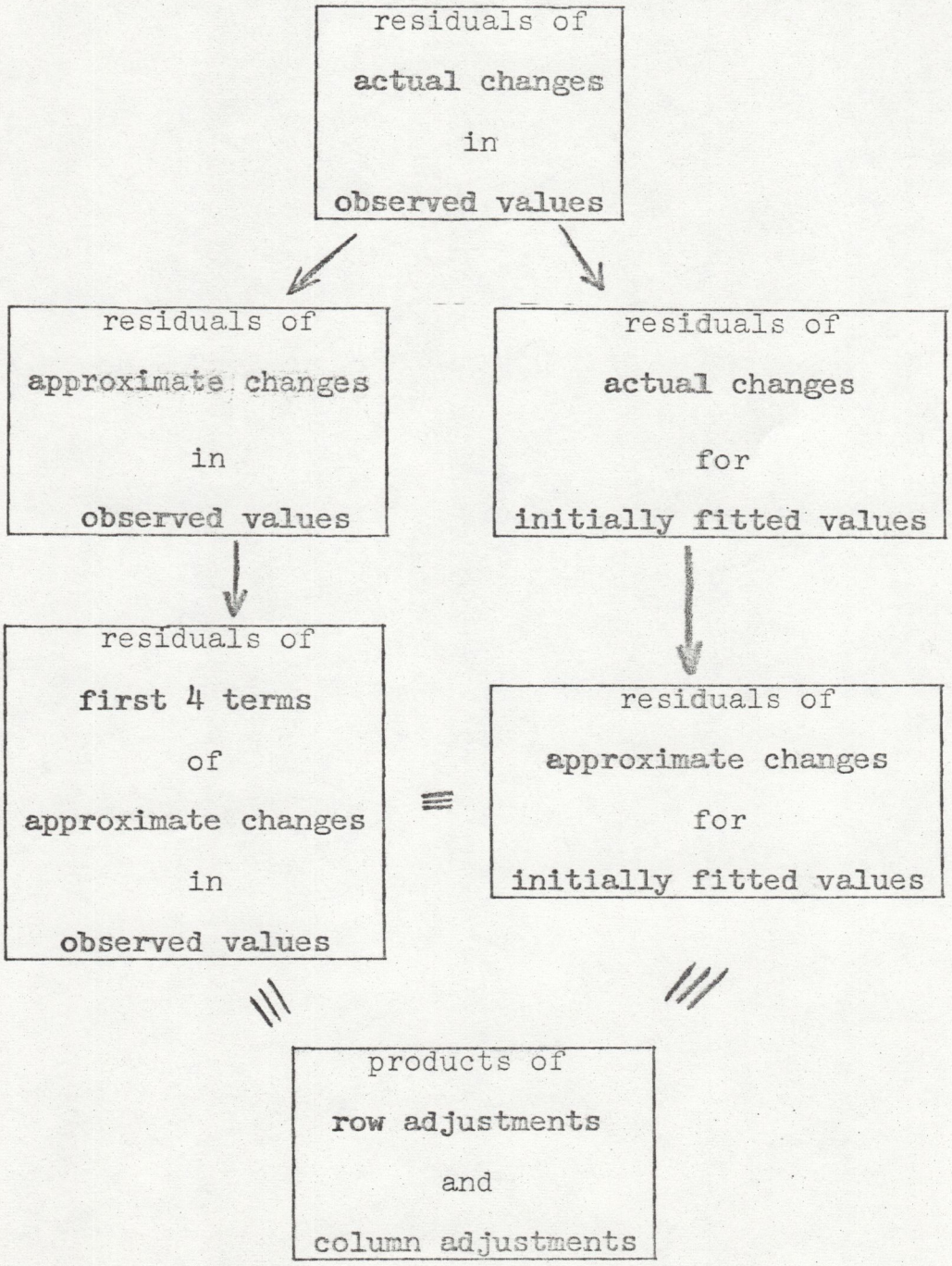


table 29 of chapter 6

The Example of table 6 Reexamined

a) DATA AND ELEMENTARY ANALYSIS, REARRANGED

				...	17	-7	0	9	...
	25	36	47	1	19	-4	0	4	1
	16	25	36	4	8	-4	0	2	4
	4	9	16	2	-8	2	0	-2	2
	1	4	9	3	-13	4	0	4	3
⋮	A	C	B		⋮	A	C	B	

b) VALUES AND ELEMENTARY ANALYSIS OF $(y-c)^2$ WITH y = OBSERVED, c = GENERAL TYPICAL

	A	C	B
1	8^2	19^2	30^2
4	$(-1)^2$	8^2	19^2
2	$(-13)^2$	$(-8)^2$	$(-1)^2$
3	$(-16)^2$	$(-13)^2$	$(-8)^2$

	A	C	B	
1	-315	3	425	245
4	-81	3	183	-52
2	87	3	-177	-52
3	69	3	-219	53
	0	-21	96	134

c) VALUES AND ELEMENTARY ANALYSIS OF $(y-c)^2$ WITH y = FITTED, c = GENERAL TYPICAL

	A	C	B
1	12^2	19^2	28^2
4	1^2	8^2	17^2
2	$(-15)^2$	$(-8)^2$	1^2
3	$(-20)^2$	$(-13)^2$	$(-4)^2$

	A	C	B	
1	-266	0	342	245
4	-112	0	144	-52
2	112	0	-144	-52
3	182	0	-234	53
	0	-49	32	165

d) PRODUCT, NOT SUM, OF ROW AND COLUMN ADJUSTMENTS FROM (a), AND THE RESULTS OF DOUBLING ALL VALUES THUS OBTAINED

	A	C	B
1	-133	0	171
4	-56	0	72
2	56	0	-72
3	91	0	-117

	A	C	B
1	-266	0	342
4	-112	0	144
2	112	0	-144
3	182	0	-234

e) ACTUAL CHANGES IN RESIDUALS ON GOING TO $\sqrt{y}$

	A	C	B
1	-4	0	4
4	-2	0	2
2	2	0	-2
3	4	0	-4

"matched" versions of each of these forms. We are allowed to change any expression, z, to A + Bz where A and B are fixed constants. Such linear coding does not, of course, alter the form of expression.

With two constants to be chosen, we can match A + Bz to y in two ways. Which two should we choose? All we can do by changing A is to move A + Bz up or down. The best that we can do with it is to make A + Bz equal to c when y = c. By changing B, we can make the changes in A + Bz for y near c close to the corresponding changes in y.

The formulas that do this are not too obvious, as shown by the formulas for the forms that include y^2, $\sqrt{y}$, log y, $-1/\sqrt{y}$, and $-1/y$, namely:

$$\frac{1}{2c} y^2 + \frac{c}{2}$$

$$2\sqrt{cy} - c$$

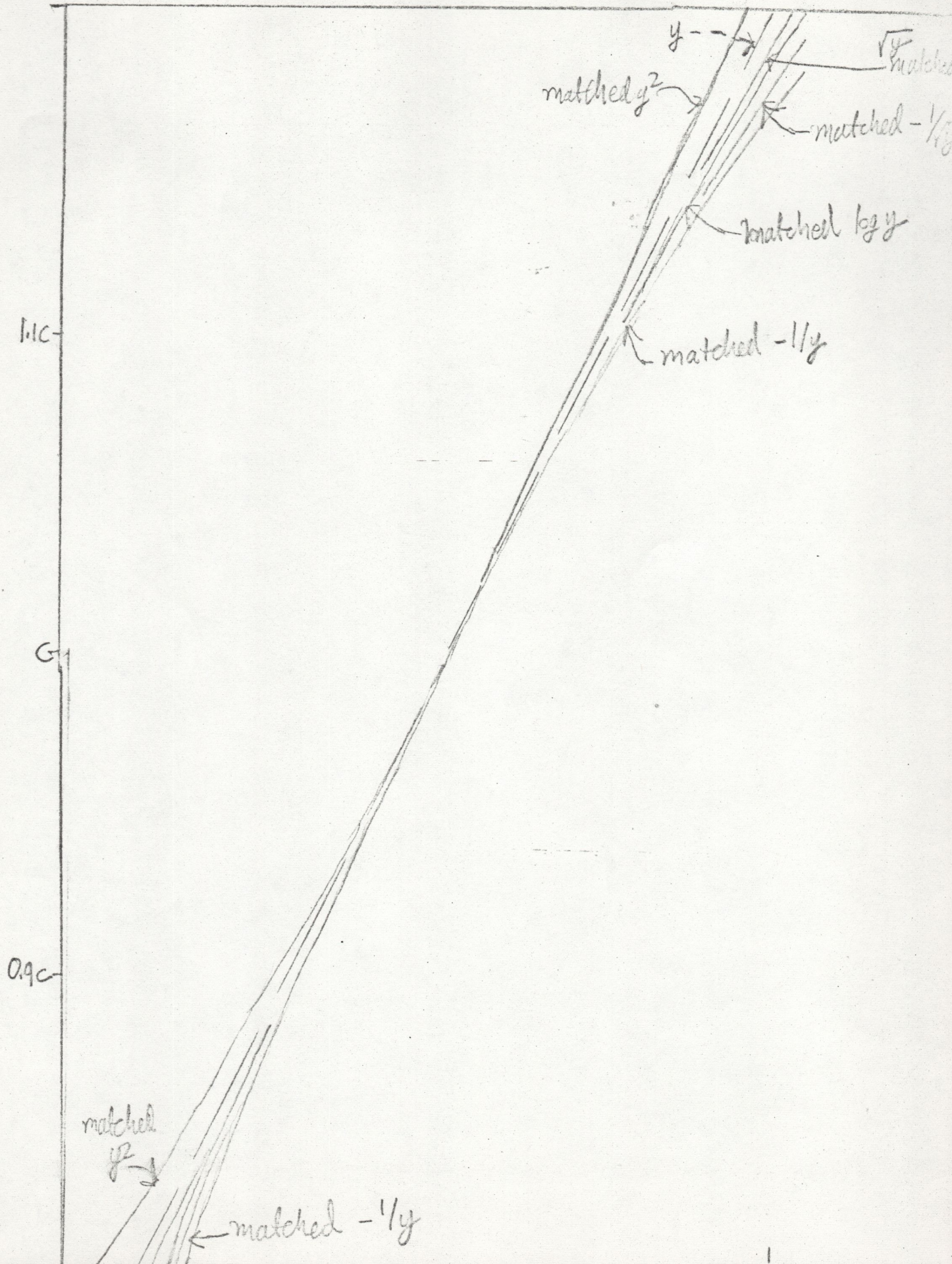
$$c \log y - c \log c + c \qquad \text{(here log} = 2.303 \log_{10}\text{)}$$

$$-2\sqrt{\frac{c^3}{y}} + 3c$$

$$-c^2/y + 2c$$

The adequacy of these matches can be judged from plot 30, which covers an interval of ±20% around c. Clearly these particular expressions are closely matched to y near c.

The matched powers compared to y near c.



When we look at this figure, we get the impression that the various matched powers deviate from y by different multiples of the same function. Over an interval somewhat longer than the interval shown in figure 27, this is reasonably so. The first approximations are set out in table 31, with a few comments.

The adequacy of these approximations can be judged from plot 32, which shows the ratio

$$\frac{\text{change according to first approximation}}{\text{actual change}}$$

for values of the c-matched power between 0.6c and 1.4c. This ratio deviates from 1.00, but stays close enough to make the first approximation an adequate tool in exploring residuals for evidence of a need for a change in form of expression.

So long as we are looking for a "loud whisper", we are ordinarily concerned with quite limited intervals for the values of the matched power. Accordingly, we can usefully look for evidence of a need for a change of form of expression by looking for changes that are proportional to

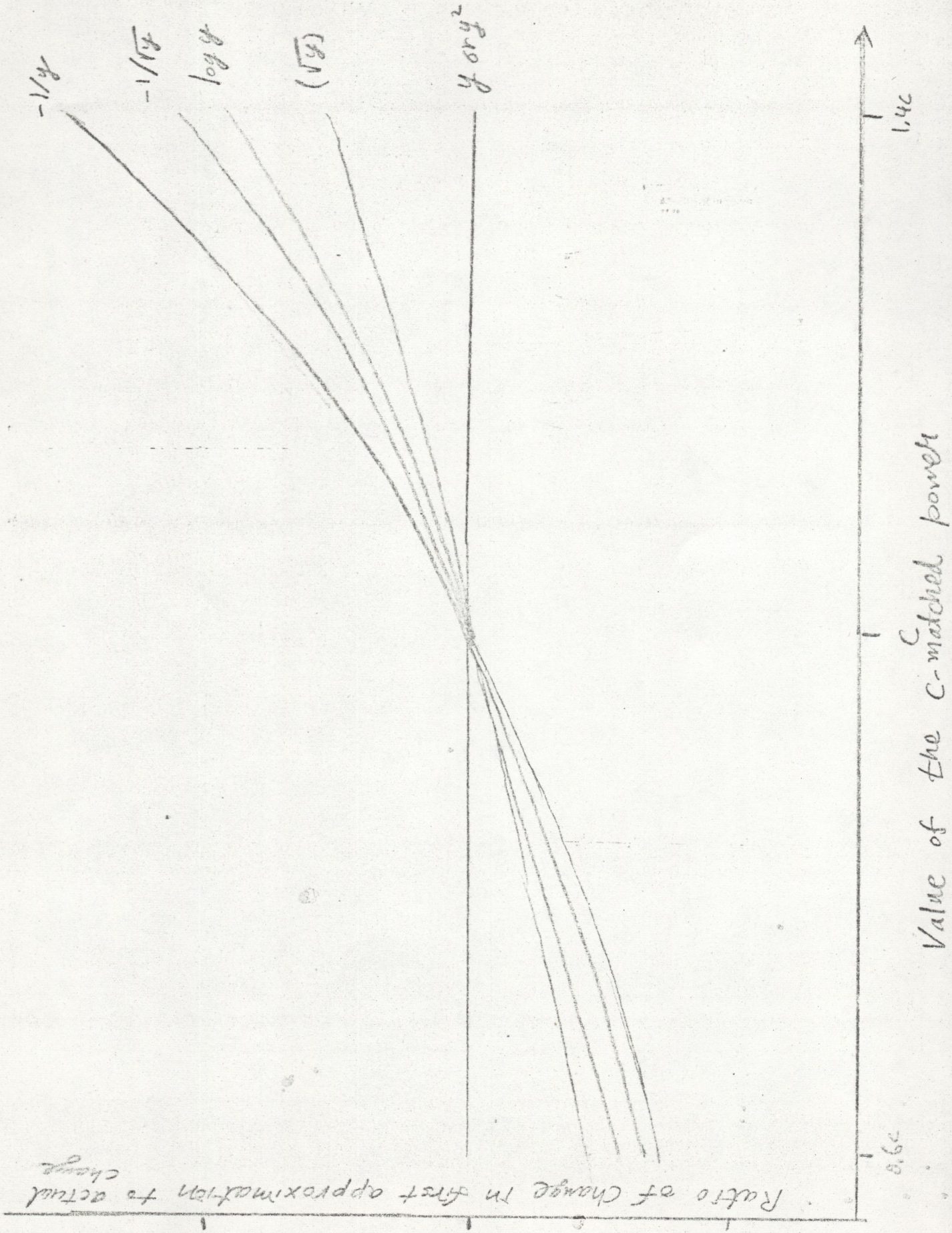
- $(y-c)^2$, when c is nearly constant.
- $(y-c)^2/c$, when c must vary (the constant of proportionality then being p-1).

table 31 of chapter 6

Forms of Expression, c-Matched Versions,
And First Approximations

Form of expression	c-matched version	First approximation	Comments
y^2	$\frac{1}{2c} y^2 + \frac{c}{2}$	$y + \frac{1}{2c} (y-c)^2$	Exact
$\sqrt{y^3}$	$\frac{2}{3}\sqrt{\frac{y^3}{c}} + \frac{c}{3}$	$y + \frac{1}{4c} (y-c)^2$	
y	y	y	
$\sqrt{y}$	$2\sqrt{cy} - c$	$y - \frac{1}{4c} (y-c)^2$	
$\log y$	$c \log y - c \log c + c$	$y - \frac{1}{2c} (y-c)^2$	log is $\log_e = 2.303 \log_{10}$
$- 1/\sqrt{y}$	$- 2\sqrt{\frac{c^3}{y}} + 3c$	$y - \frac{3}{4c} (y-c)^2$	
$- 1/y$	$- \frac{c^2}{y} + 2c$	$y - \frac{1}{c} (y-c)^2$	$y - \frac{1}{y} (y-c)^2$ is exact
y^p or $- /y^{-p}$	$\frac{1}{p} y^p + \frac{p-1}{p} c$	$y + \frac{(p-1)}{c} (y-c)^2$	General form

Ratio of changes, (first approximation)/(actual)



6T. <u>Why is (row adjustment)x(column adjustment) its own residual?</u>

We claimed above, in section 6R, that a table of the form (row adjustment)x(column adjustment), when subjected to elementary analysis is its own residual.

If we are using midmean differences in our standard way, the midmean of all row adjustments will have been made zero. If the column adjustments for the ith and jth columns are c_i and c_j, the differences for those columns take the form

$$(c_i - c_j)(\text{row adjustments})$$

whose midmean will be zero, since $c_i - c_j$ is a constant. Thus all entries in the square combining table will be zero and the fit by columns will be zero.

Interchanging rows and columns, we see that the same result holds for the fit by rows. Since nothing is fitted, the residuals are the original table.

If our analysis is based on mean differences, we routinely make the mean -- not the midmean -- of the row adjustments vanish. Hence the mean of

$$(c_i - c_j)(\text{row adjustment})$$

vanishes again, making every entry of the square combining table vanish. Again the table will be its own residual.

((If we had adopted some mixed-up standard pattern of analysis, such as analysis by midmean differences with the

means -- not midmeans -- of the stripe adjustments made zero, we would not have had such a handy result.))